Traffic Accidents Analyzer Using Big Data

Eyad Abdullah
Information System Department
College of Computer and Information Sciences
King Saud University, Riyadh, Saudi Arabia
sfbit01_beta@live.com

Ahmed Emam
Information System Department
College of Computer and Information Sciences
King Saud University, Riyadh, Saudi Arabia
Menoufia University, Menoufia, Egypt
aemam@kus.edu.sa

***Abstract*— Traffic accidents are serious issues, which can possibly cause disabilities, injuries and even fatalities. In order to decrease the number of accidents, we need understand and analyze the traffic accidents dataset. Almost every day, someone has sufferes from traffic accidents in one way or another, such as the traffic being slowed down due to an accident or collision in the same road, which may leave one or more lanes unavailable. Big Data ecosystem has the ability to store, manipulate, analyse, and mine large traffic accident datasets and can drive knowledge creation that can help decision makers to reduce the number of accidents. Massive real traffic datasets from New York's traffic collisions dataset is used as source of data for the developed application. The developed application consists of several functions and web services to analyze and visualize the major traffic accident information. The developed application stores the massive traffic data on Hadoop with a parallel computing framework for diverse and mine based Map-Reduce technique, which then uses Web services interface to support developed mining application.**

Index Terms: Big Data, Road traffic accidents, Data Mining, Hadoop and Map-Reducer.

I. Introduction

Traffic accidents have a great economic impact due to their cause of injuries and fatalities. Nowadays, many researchers give much attention to determining common factors that significantly affect traffic accidents and analysis. There are several approaches that researchers have applied to investigate this problem, such as Artificial Neural Network, Data Mining, Logic Formulation, and Fuzzy ART Maps. To obtain the greatest possible accident reduction effects, with limited budgetary resources, it is important that measures and analysis should be based on scientific and objective research of the causes of accidents [1]. The data mining researcher mandated to develop a method for the analysis of the causes of road traffic accidents based on data mining that analyzed the related attributes and causes of road traffic accidents. Traffic Accidents Analyzer is a software written in C# programming language with the use of SSH library [10]. It analyzes traffic accidents to give the user general knowledge of the accidents which will aid the process of identifying problems and providing solutions. The analysis of traffic accidents is done through MySQL aggregation functions (such as SUM and COUNT), and will also use some data mining techniques, using Mahout and Hadoop in its analysis. The application itself doesn't do any processing of data, it is just sends and receives commands from the server; the server will handle all the processing. The software uses the capabilities of Hadoop, Mahout and MySQL. Apache Hadoop is a framework written in Java for distributed storage and distributed processing of very large data [7]. Apache Mahout is a project by Apache to produce free implementations of scalable machine learning and data mining for large data; it needs Hadoop in order to work. Mahout distributes the work into many nodes in the network to speed up the processing, and it supports many data mining techniques and algorithms such as clustering, classification, Collaborative Filtering and many others. MySQL is an open source database management system that can support huge amounts of data [9]. Since the software is using the capabilities of these technologies, it will have a good scalability in which it can handle huge amount of accidents data. Hadoop, Mahout and MySQL are residing in the same server and under the same Operating system of CentOS (Linux distribution). The software will connect to CentOS using SSH network protocol in order to use Hadoop, Mahout and MySQL. The application has been written as multithreaded, in which one thread is responsible for handling the user interface and receiving inputs from the user and the other thread will handle the communication between the user (client) and the server, so that in case the application is waiting for the results from the server, it will not hang (unlike single threaded), and will remain responsive at all times. Also, the software will tell the user the progress of the executing operation and allow the user to cancel any operation at any time. This is very important in case the operation takes minutes or hours to complete because of the large size of the data. Because of the multithreaded architecture of the software, it will have a good availability.

II. Pervious Research Stuides

In the article by Chong [1], he stated that applying data mining techniques to model traffic accident data records can help to understand the characteristics of drivers' behavior, roadway condition and weather condition that were causally connected with different injury severity. At the same time, it will help decision makers to formulate better traffic safety control policies. The author analyzed the GES automobile accident data from 1995 to 2000 and investigated the performance of neural network, decision tree, support vector machines and a hybrid decision tree with neural network techniques for predicting

978-1-4673-9795-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCI.2015.187

drivers' injury severity in traffic accidents. The author developed a set of experiments with a predicated model that classifed fatal and non-fatal injury. The only drawback of this study was the used dataset didn't provide enough information on the actual speed,and other information such as road condition. Yang et al [2] used the neural network approach to detect safer driving patterns that have less chances of causing death and injury when a car crash occurs. In the article by Roh [3], he illustrated how statistical methods based on directed graphs, constructed over data for the recent period, may be useful in modeling traffic fatalities by comparing models specified and using directed graphs to a model, based on out-of-sample forecasts. In the research article by Rui [4], the author stated that the road factors that play a vital role in traffic accident are the linear, slope, linear combination and road surface conditions. He also mentioned the vehicle factors, such as the quality of technical condition of the vehicles, the motor factor, steering, braking, driving and electric factors. The author extracted 100 datasets from the traffic accidents database of a city, and then analysed the road traffic accident data using statistical data analysis that support the road traffic safety related research method. In the research article by Shi [5], he stated that understanding traffic flow trend on highways under collisions is significant to reducing the impact of traffic accidents. The article proposed a method to constructing time-series data using traffic flow data when accidents happened. To avoid the defect of not considering the linear drift in the time domain between two sequences, DFT was carried out to extract features from original time-series data. The traffic flow trend could then be well understood by clustering analysis. The case study, using real data, in Harbin showed the feasibleness of the proposed method on Real datasets extracted from records collected on Beijing-Harbin Expressway (G1) between Harbin and Lalinhe from January 2010 to July 2011. Traffic flow data transformed into time series data by the Cell Transmission Model (CTM) method and used Discrete Fourier Transform (DFT) to extract the most important features, then apply the clustering analysis method to understand the traffic flow trend on highways. Yu [6] believes that traffic information mining is a typical "Big Data" application with traffic data exceeding 1.5PB, and the storage capability over 5PB, and about 800GB of data is produced each day in Beijing. The author stated that traffic data mining will utilize history traffic data and discover transportation mode while applying common association rules through data mining techniques. The author developed parallel functions that process distributed and blocked traffic data based on Map-Reduce framework, which contain Accident Detection and Traffic Trend Prediction. The developed cloud services are based on a RTIC-C (4 layers to support distributed storage, parallel computing and customized services for massive traffic data) system and can handle massive traffic data sets including floating car, mobile phone, and bus data set. The developed traffic data mining was based on cloud computing technique and built with a massive traffic data storage base on Hadoop with a parallel computing framework for diverse kinds of mining applications based on Map-Reduce mechanism, and a restful Web services interface to support third-party mining applications.

III. Proposed Architecture

The proposed application will work on the operational database (MySQL), and in case the application is requested to apply data mining technique, it will extract the data required from the operational database and pass that data to Hadoop, Mahout is then used to analyze the extracted data. The operational database that has been used is New York's traffic collisions dataset [11] which is used to simulate the operational database. Selected attributes from Dataset

Attribute Name	Attribute Description
Time	The time of the accident in (24 hour)
Borough	The administrative division
Latitude	The latitude of the accident
Longitude	The longitude of the accident
On_Street_Name	The street of the accident
Persons_Injured	Total number of people injured
Persons_Killed	Total number of people killer
Cause	Causes for that accident.
Vehicle_Types	Vehicle types for that accident it may contain one type or multiple types of vehicles separated by comma

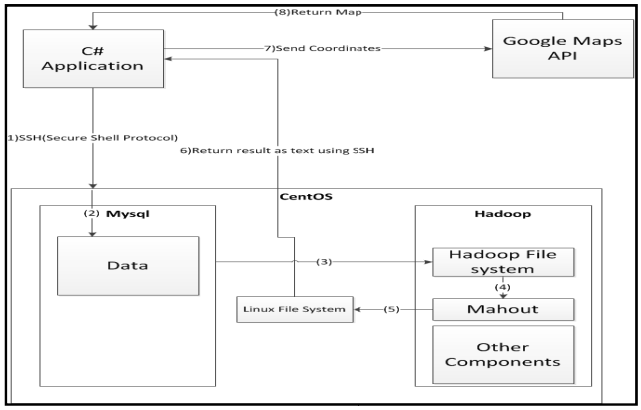

Figure 1: Architecture of the proposed Application.

Unfortunately, the dataset includes many missing and redundant values, and it is not normalized. After the data is uploaded to MySQL, the records that have the missing values for the important attributes that cannot be fixed will be deleted. The attributes that describe the same thing are grouped into one attribute and the values are separated by a comma. See Table 1 for the attributes that are used in the analysis of data (the attributes that are not used are not included in the table). Also, see Fig. 1 for the architecture of the project.

IV. Proposed Application Operation

When the user runs the application, the application will connect to the server using SSH and a session is established with

CentOS. Fig. 2 shows the main interface for the application. The application can provide 6 different analysis functions. The analysis results are converted into tabular view, chart view and map view (depending on the function). The tabular view shows the result as a table. The Chart view shows the result as a Chart in which the x-axis contains the common values and the y-axis contains the frequency for these values. The map view extracts the GPS coordinates from the results and then plots these coordinates into a map (using Google Maps API). The user has the choice to analyze traffic data for a specific administrative division (borough) or the administrative divisions as a whole. We will discuss these functions in a good detail.

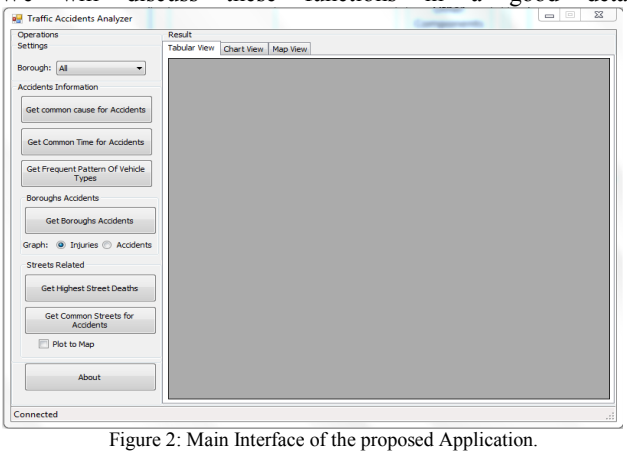

Figure 2: Main Interface of the proposed Application.

Based on the previous studies, the proposed application can analyze the following typs of data: Accident Information, Brough Accidents, and Street Related. Each direction will consist of several function and each function will perform achieve specific task.

A. Common cause for Accidents Function

This *function "Get common cause for* Accidents" will execute the SQL query, which will apply the aggregate function COUNT in MySQL to get the common causes for accidents and sort the results in descending order based on the number of accidents.

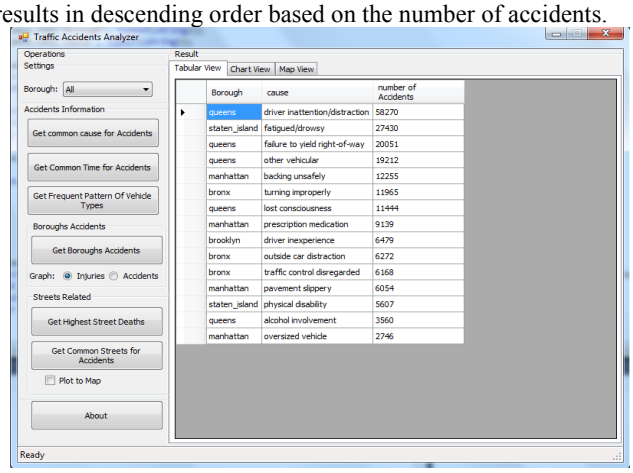

Figure 3: Tabular view of "Get common cause for Accidents

The final result will be formatted into a table that has three columns: Borough, Cause and Number of Accidents as shown in Fig. 3. At the same time, Fig. 4 shows the distribution of the major accidences causes for the sleeted period of time.

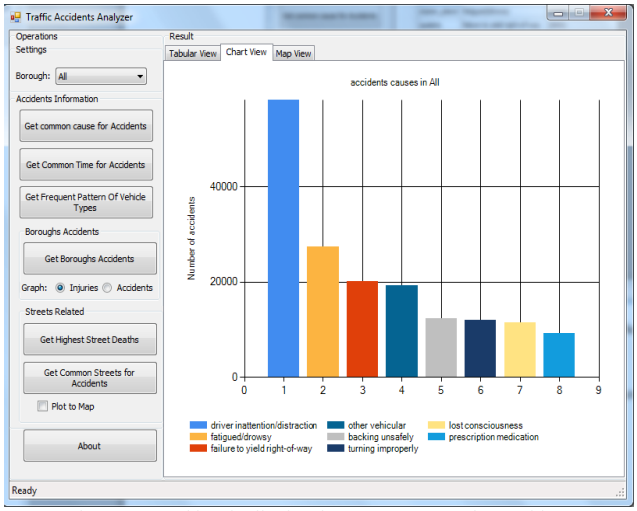

Figure 4: Graphic Distribution for common cause for Accidents.

B. Common Time for Accidents Function

This function goal is to "Get Common Time for Accidents" function. It will execute the SQL query which will apply the aggregate function COUNT in MySQL, to get the common time for accidents and sort the results in descending order based on the time of accidents as shown in Fig.5.

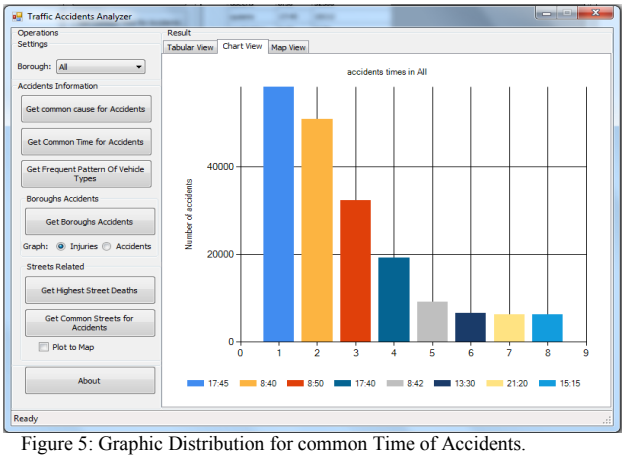

Figure 5: Graphic Distribution for common Time of Accidents.

C. Frequent Vehicle Type particpate in Accidents Function

This function applies the data mining technique FPgrowth, using Mahout to extract the frequent patterns of vehicle types

that appear in accidents, and then sort the result in descending order. The result of this function is a table that has two attributes: Vehicle Types Patterns and Accidents (number of accidents). The algorithm for the designated function will work as follows:

1. Remove the old Mahout Result files from Hadoop.
2. Remove the old input data and the results files from CentOS.
3. Remove the old input data and the results files from Hadoop.
4. Execute a SQL query to count the number of records in the MySQL database (to calculate support for step 8).
5. Execute SQL query to export values of the attribute 'vehicle_types' only into a CSV file.
6. Execute a Linux's "sed" command to remove the quotes from the CSV file.
7. Upload the file into Hadoop. See Figure 7
8. Apply FPgrowth using Mahout with the method "mapreduce" with support of 1% (this step may take few minutes).
9. Download the results from Hadoop file system into CentOS file system.
10. Use the Linux's command "cat" to show the results.
11. Parse the result into tabular view and chart view.

The Mahout raw results shown in Fig. 6 and the graphical representation of the Get Frequent Pattern of Vehicle Types Function presented in Fig. 7.

```
Input Path: patterns/frequentpatterns/part-r-00000
Key class: class org.apache.hadoop.io.Text Value Class: class org.apache.mahout.
Max Items to dump: 50
Key: SUV: Value: ([SUV],81192), ([passenger_vehicle, SUV],39750), ([SUV, unknown
Key: bicycle: Value: ([bicycle],5224), ([passenger_vehicle, bicycle],2515)
Key: bus: Value: ([bus],7712), ([passenger_vehicle, bus],3543)
Key: large_vehicle: Value: ([large_vehicle],8132), ([passenger_vehicle, large_ve
Key: livery_vehicle: Value: ([livery_vehicle],4981)
Key: other: Value: ([other],12045), ([passenger_vehicle, other],5383), ([SUV, ot
Key: passenger_vehicle: Value: ([passenger_vehicle],159087), ([passenger_vehicle
Key: pick-up_truck: Value: ([pick-up_truck],6314), ([passenger_vehicle, pick-up_
Key: small_vehicle: Value: ([small_vehicle],7605), ([passenger_vehicle, small_ve
Key: taxi: Value: ([taxi],16150), ([passenger_vehicle, taxi],6533), ([SUV, taxi]
Key: unknown: Value: ([unknown],29151), ([passenger_vehicle, unknown],16598), ([
Key: van: Value: ([van],14037), ([passenger_vehicle, van],6251), ([SUV, van],344
Count: 12
```

Figure 6: Mahout Raw Results.

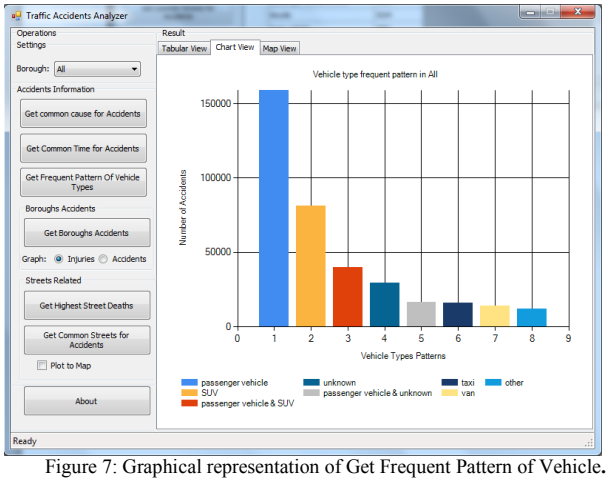

Figure 7: Graphical representation of Get Frequent Pattern of Vehicle.

D. Districts Accidents Function

This function executes the SQL query, which uses the aggregation functions COUNT and SUM, to count the number of accidents for each borough, it will then sum up the total injuries and deaths for that borough. It gives two options for creating the charts, one for injuries one for deaths. See Fig. 8 for more details.

E. *Get Highest Street Deaths" Function*

This function executes the SQL query that uses the aggregation function SUM. The result contains the street names and the total deaths which occurred in each street, in descending order. The final results from deploying this function can be represented either graphically in a chart view or a map view as shown in Fig. 9.

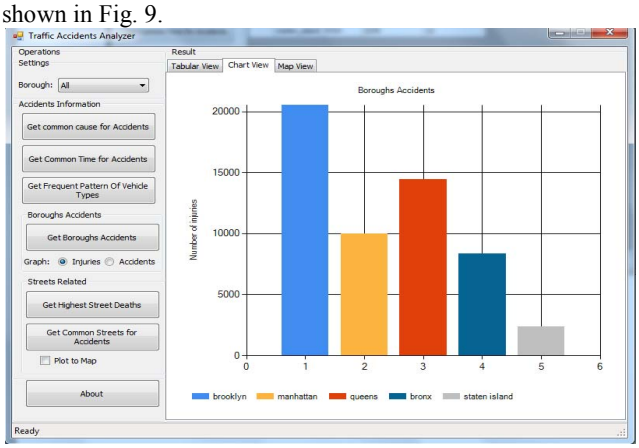

Figure 8: Chart view for Get Boroughs Accidents function.

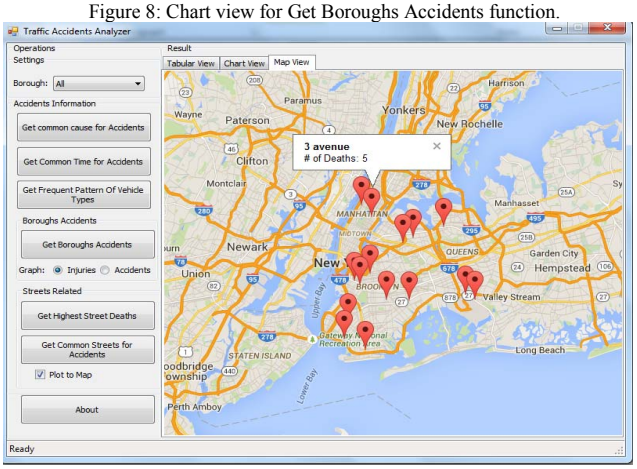

Figure 9: Map view for Get Highest Street Deaths function.

F. Common Streets for Accidents Function

This function executes a SQL query, which uses the aggregation function COUNT, to count the number of streets.

The result contains the street names and the number of accidents that happened in each street, sorted in descending order as shown in Fig. 10.

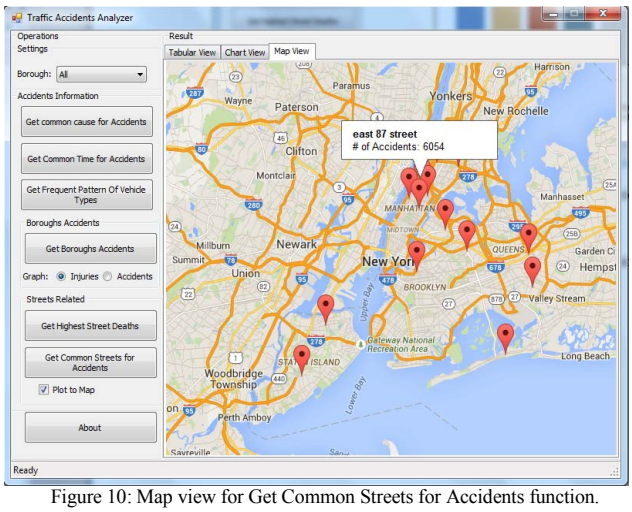

Figure 10: Map view for Get Common Streets for Accidents function.

V. Traffic Accidents Analyzer as Web Services

An enhanced version of the proposed application has been created as web services which can be invoked by many clients such PHP,ASP, mobile phones or any device or application that can send POST requests, or using SOAP, which is a protocol for communicating with web services. The enhanced version allows the capabilities of Hadoop and Mahout to be utilized online and remotely. Fig. 11 shows the version of enhanced architecture for proposed application based on service.

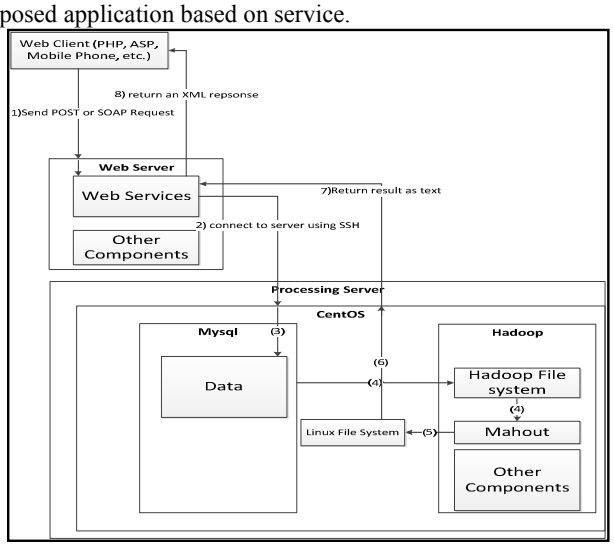

Figure 11: Service based for application architecture.

Figure 12 represent screenshot of web pages (HTML) that use JavaScript (AJAX) to connect to the PHP server, which in turn connects to the web service, and is accessed by a Computer and an iPhone.

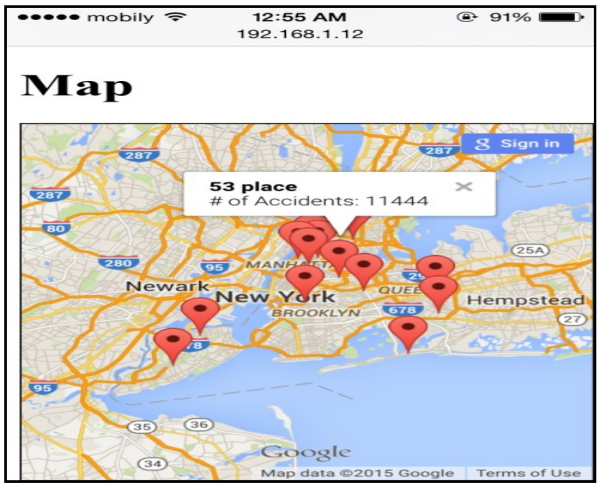

Figure 12: Map view for Get Common Streets for Accidents function

VI. Conclusion

Traffic accidents are serious issues, which can possibly cause disabilities, injuries and even fatalities. In order to decrease the number of accidents, we need to understand and analyze them. Since traffic accidents data is generated almost daily,and the data size will increase very fast in all dimensions of data, the urgent need for an application that handles this growth very well and analyzes the traffic accident has become essential. Nowadays, the Big Data ecosystem has the ability to store, manipulate, analyze, and mine large traffic accident datasets and drive knowledge creation that can help decision makers to reduce the number of accidents. A few researchers give much attention and focus to analyzing the large traffic accident dataset using big data approach. This study presents a very important application tool for using big data for storing, integrating, and analyzing the traffic accidents using Mahout Data Mining as a part of big data ecosystem. Very large and real traffic datasets from New York's traffic collisions dataset is used as source of data for the developed application. The developed application consists of several functions and web services to analyze and visualize the major traffic accident information. The developed application stores the massive traffic data on Hadoop with a parallel computing framework for processing and mining based on Map-Reduce technique, then uses Web services interface to support developed mining application.

REFERENCES

[1] Miao Chong, Ajith Abraham2 and Marcin Paprzycki1, "Traffic Accident Analysis Using Machine Learning Paradigms", Informatica Journal Vol. 29, pp. 89–98, 2005.

[2] Yang, W.T., Chen, H. C., & Brown, D. B.,Detecting Safer Driving Patterns By A Neural Network Approach. ANNIE '99 for the Proceedings of Smart Engineering System Design Neural Network, Evolutionary Programming, Complex Systems and Data Mining, Vol. 9, pp. 839-844, November 1999.

[3] Roh J.W., Bessler D.A. and Gilbert R.F.**,** Traffic fatalities, Peltzman's model, and directed graphs**,** Accident Analysis & Prevention, Vol. 31, Issues 1-2, pp. 55-61, 1998.

[4] Rui Tian, Zhaosheng Yang and Maolei Zhang, "Method of Road Traffic Accidents Causes Analysis Based on Data Mining", 2010 International Conference on Computational Intelligence and Software Engineering (CiSE), pp. 1 – 4, 10-12 Dec. 2010, DOI:10.1109/CISE.2010.5677030, 2010.

[5] An Shi, Zhang Tao, Zhang Xinming, and Wang Jian, "Evolution of Traffic Flow Analysis under Accidents on Highways Using Temporal Data Mining", 2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications, DOI 10.1109/ISDEA.2014.109, pp.454-457, 2014.

[6] Jianjun Yu, Fuchun Jiang, and Tongyu Zhu, "RTIC-C: A Big Data System for Massive Traffic Information Mining", 2013 International Conference on Cloud Computing and Big Data, DOI 10.1109/CLOUDCOM-ASIA.2013.9, pp.395-402, 2014.

[7] http://en.wikipedia.org/wiki/Apache_Hadoop last visisted July 2015.

[8] http://en.wikipedia.org/wiki/Apache_Mahout, retrieved May 6, 2015.

[9] http://www.tesora.com/myth-4-mysql-cannot-handle-large-volumes-data-especially-queries-joins-and-aggregations-i-must/ retrieved June 10, 2015.

[10] https://sshnet.codeplex.com/ retrieved May 10, 2015.

[11] https://data.cityofnewyork.us/NYC-BigApps/NYPD-Motor-Vehicle-Collisions/h9gi-nx95. retrieved July 6, 2015.